

Improved Echo State Network (ESN) for the Prediction of Network Traffic

Dezhong Ye¹, Haibing Lv¹, Yong Jiang¹, Zhongzheng Wu¹, Qiuxia Bao^{2,3}, Yun Gao^{2,3},
Ruo Chen Huang^{2,3}
{ye.dezhong, lv.haibing2, jiang.yong}@zte.com.cn

ZTE CORPORATION, China, 210012¹, College of Telecommunications and Information Engineering,
Nanjing University of Posts and Telecommunications, Nanjing, China, 210003², National Engineering
Research Center of Communications and Networking (NUPT), China, 210003³

Abstract. With the development of computer and network technologies, the network traffic has come to an explosive growth. Considering the non-linearity of network data, we choose echo state network (ESN) to predict the data. ESN is a novel kind of recurrent neural networks, where a reservoir is generated randomly and only output weight matrix is adaptable. In our paper, we propose an improved network of ESN considering its weakness on selection of matrix initialization and activation function. The improved network defines the scope of matrix initialization and replaces the activation function of middle layer with wavelet function. And result shows that our improved network is more effective compared to original ESN. Our evaluation index is normalized mean square error (NMSE), and it drops from 0.7435 to 0.5852 by making improvements.

Keywords: Echo state network (ESN), Weight matrix, Wavelet function.

1 Introduction

Nowadays, the number of Internet users is increasing dramatically, and the scope of Internet services is also becoming wider. China Internet Network Information Center(CNNIC) shows that the number of netizens in China had reached 772 million by December 2017 [1]. People have a growing need for network traffic. Therefore, research on the new network with excellent ability of network perception and intelligent decision has become a hot spot in the industry [2] [3].

As one part of network perception, traffic prediction can greatly improve the quality of service and the performance of network. On the other hand, it can bring network operators more optimized service policies and better distribution of network resource. Ultimately, users can enjoy a higher quality of experience. With in-depth research on neural network, researchers begin to use neural network to predict nonlinear time series. The survey found that BP neural network and its improved algorithms are widely used. However, BP neural network uses gradient descent learning algorithm, which limits the size of the network. At the same time, it generally needs long training time. The result is easy to fall into the local optimal solution. Later, researchers find that dynamic neural networks have better predictive effect on nonlinear time series. Thus, a new recurrent neural network, echo state network (ESN), is proposed. ESN discards gradient learning algorithm and uses a simple and effective single training algorithm, which greatly shortened the network training time. More importantly, the

network predicts well. Numerous experiments show that ESN has better adaptability to nonlinear time series.

The rest of this paper is organized as follows. Section 2 describes previous related work about prediction of network traffic. In Section 3, we make an analysis of the data acquired from telecom operator. In Section 4, we introduce the structure of echo state network (ESN) and improvement of model. Finally, we give simulation results and conclusion in section 6.

2 Related work

Both short-term and long-term prediction of network traffic are feasible directions for research. With historical traffic data and suitable method, we can have a general idea of the size of future traffic. With the prediction, operators can adjust the distribution of network resource in time. Also, equipment manufacturers can replace network devices with insufficient load capacity in advance.

In recent years, there have been many technologies applied in the field of traffic prediction. Pang et al [4] established an adaptive fuzzy traffic predictor based on the theory of fuzzy system [5]. They used nearest neighborhood clustering learning algorithm to present the fuzzy traffic predictor, and the results showed that the predictor was accurate and flexible applied in ATM networks. Sang et al [6] predicted network traffic using two stationary traffic models, including the Auto-Regressive Moving Average (ARMA) [7] model and the Markov-Modulated Poisson Process (MMPP) [8] model. These two models mainly use mathematical analysis and experiments to do the predictability analysis of network traffic. Besides, wavelet analysis has been one of the most effective methods to deal with non-stationary time series for traffic prediction. Wang et al [9] proposed a novel method of combining wavelet and RLS to forecast the Internet traffic and results showed this method achieved extraordinary accuracy compared to other models. As we all know, Artificial Intelligence (AI) has been applied in many areas widely, such as expert system, fuzzy reasoning and fuzzy neural. Specially, there are many techniques about neural networks (NNs) used in the field of prediction. Neural networks are suitable for learning more complex nonlinear relationships. Moreover, it's data-driven and continuously learned, and does not require extensive mathematical mathematics of network traffic. Echo state network (ESN) is a novel kind of recurrent neural networks, where a reservoir is generated randomly and only readout layer is adaptable [10]. With the advent of video services, requirements on traffic transmission are also improved. [11] and [12] propose an energy-efficient content delivery system to improve the ability of traffic transmission. And [13] gives an effective calculation method to transfer information with the least resource.

Compared with other types of neural networks, echo state network (ESN) is equipped with simple network structure and prediction of exact nonlinear chaotic time series. Besides, this network has good generalization ability for nonlinear time system. For these reasons, we propose the ESN model to establish network traffic predictor in this paper.

3 Data analysis

3.1 Data processing

Here, we use the real data from ZTE CORPORATION which lasted from March 2 to April 6, 2017. Traffic data in the device was recorded every fifteen minutes. Firstly, we need

to remove outliers in the data caused by machine failures. To observe the changes in network traffic of the same port, we pick out the records with the same noid. Then, we concatenate the selected records into a continuous time data.

Figure 1 gives a visual description of part of the traffic data. One record in the dataset represents a time period of fifteen minutes, so one day at a total of 96 records. From the figure, we can see that data waveform has a regular pattern and the length of the circle is exactly about one hundred. Therefore, we reasonably guess that traffic data has a certain regularity in the dimension of days. Also, it is obvious that there are some special points in the data curve. We guess that some things may happen in these days, which is exactly the difficulty of our prediction work.

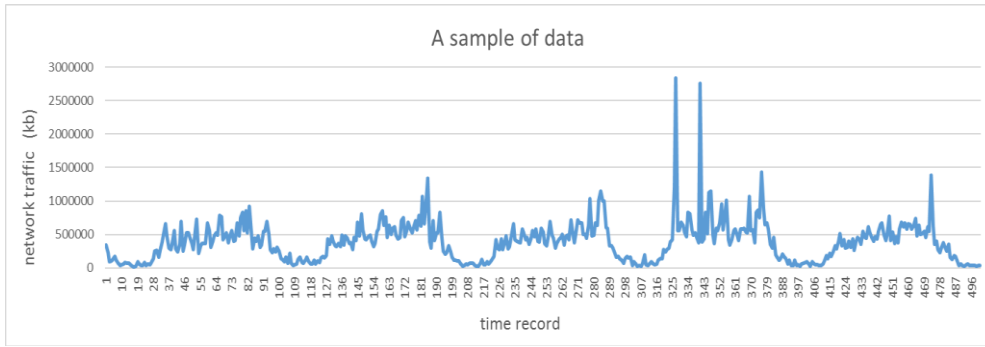


Fig. 1. A sample of data.

3.2 Result assessment

In this paper, we use normalized mean square error (NMSE) to indicate the prediction ability of ESN. NMSE is calculated as follows

$$R_{NMSE} = \frac{\overline{\|\hat{y}(n) - y(n)\|^2}}{\overline{\|y(n) - y(n)\|^2}} \quad (1),$$

where $\hat{y}(n)$ is the predicted output and $y(n)$ is the target output. $\|\cdot\|$ represents Euclidean norm. $\overline{\cdot}$ represents the average of \cdot .

The reasonable value range of NMSE is 0 to 1. A smaller result of NMSE means that the predicted value is more closer to the real value. If the result is more than 1, the model prediction has little value.

4 Echo state network (ESN)

4.1 Basic ESN

Network traffic data is non-linear time series, and it has the characteristics of nonlinearity, autocorrelation, long relevance and suddenness. As neural network prediction algorithm has excellent ability of self-learning and nonlinear approximation, it adapts well to nonlinear time series.

Echo state network is a kind of dynamic recurrent neural network. It has three layers, called input layer, middle layer and output layer. Different from the traditional neural network,

the middle layer of ESN is usually called dynamic reservoir. Dynamic Reservoir contains a large number of sparsely connected neurons, which mimics the working principle of the human brain neurons. They accept information from the input layer just as the human brain neurons receive stimuli from the outside world [14].

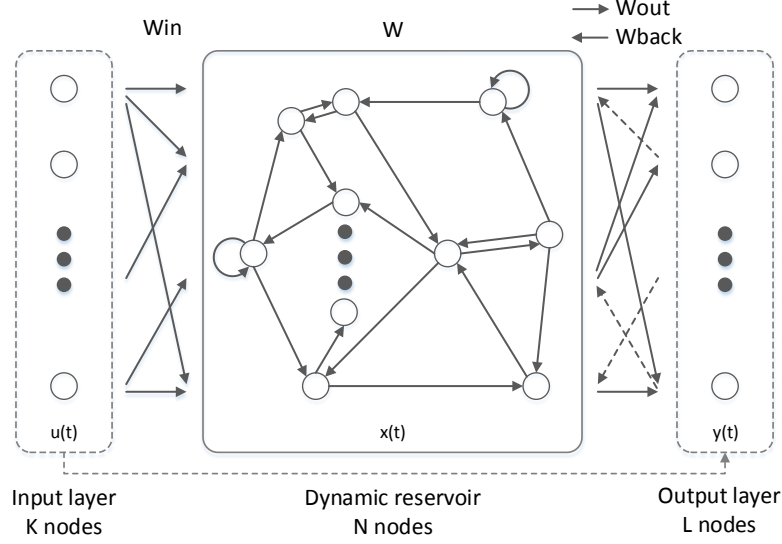


Fig. 2. Architecture of a based Echo State Network.

Figure 2 shows the architecture of a original Echo State Network. It is noted that the neurons in dynamic reservoir are not completely interconnected. These neurons can record the operating status of the system and act as a short-term memory storage. Eq.(2) describes the state of nodes in ESN.

$$\begin{aligned} u(t) &= [u_1(t), u_2(t), \dots, u_K(t)]^T. \\ x(t) &= [x_1(t), x_2(t), \dots, x_N(t)]^T. \\ y(t) &= [y_1(t), y_2(t), \dots, y_L(t)]^T. \end{aligned} \quad (2),$$

where $u(t)$, $x(t)$, $y(t)$ represent the state of the input layer, dynamic reservoir, and output layer at time t separately. Their dimensions are K , N and L respectively.

As we can see in Fig.2, there is information transmission and feedback between the three layers. This transfer relationship is reflected in the weight matrix W_{in} , W , W_{out} , W_{back} . In order to simplify the network operation, we do not consider W_{back} here. W_{in} is the matrix between input layer and reservoir, with dimension of $N * K$. And W is the matrix connecting the present and the next reservoir, with dimension of $N * N$. Then W_{out} reflects the relationship between input, reservoir and output, with dimension of $(K + N + L) * L$ [15].

W_{in} and W are randomly initialized and will not change in the following steps. The only parameter we need to adjust is W_{out} . Only when the spectral radius of W is less than 1, it can keep the ability of short-term memory. Therefore, we define W by Eq.(3).

$$W = m \cdot (W' / |\rho_{max}|). \quad (3),$$

where W' is the initial matrix. ρ_{max} is the spectral radius of W' . m is the scale parameter with the range of $[0,1]$.

Training step.

There are two most significant equations in ESN. One is the state update equation described in Eq.(4).

$$x(t + 1) = f(W_{in} \cdot u(t + 1) + W \cdot x(t)). \quad (4),$$

where f is the neuron activation function. We generally use tanh function. $x(t)$ is the state of reservoir at time t and $x(t + 1)$ is the state of the next moment. $u(t + 1)$ is the state of input at time $t + 1$. Observing Eq.(4), we can get all the state of $x(t)$ with the known W_{in} and W . Then we can get the state matrix M and output sequence T with Eq.(5).

$$M = (u^T, x^T, y^T) , \quad T = y^T . \quad (5),$$

As defined,

$$M \cdot W_{out} = T . \quad (6),$$

Then we can get W_{out} with simple linear regression. And the training process of ESN is completed.

Testing step.

Next, we can do the prediction with the output state equation. This is exactly the other significant equation described by Eq.(7).

$$\hat{y}(t + 1) = f^{out} \cdot (W_{out} \cdot (u(t + 1), x(t + 1), y(t))) . \quad (7),$$

where \hat{y} is the predicted output, f^{out} is the neuron activation function of output layer. We generally use identity function. And the testing process of ESN is completed.

The salient benefits of ESN are also the conciseness and effectiveness of the training and testing process.

The advantages of ESN are the excellent ability of prediction of non-linear time series and simple training process. However, randomness in the training process may reduce the prediction accuracy of the network. Also, the unity of memory function limits the memory capacity of the reservoir. Thus, we bring the following two effective improvements to ESN by consulting the literatures. And results indicate the effectiveness.

4.2 Weight matrix

In Eq.(4), we introduce the state update equation. The original activation function is tanh, which is showed in Fig.3. The area where the function value is close to plus or minus one is usually called the saturated area, and the other area is called the activation area. As we can see in Fig.3, curve between the saturated area is relatively flat. Then the nonlinear fitting effect of the model will be poor. Also, according to Eq.(4), variables in the saturated region may cause singular solution.

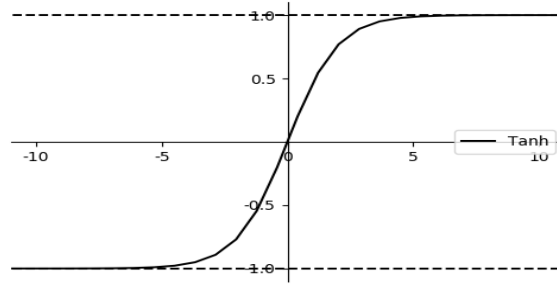


Fig. 3. tanh function.

In the based ESN, W_{in} and W are randomly initialized. W_{in} and W obey homogeneous uniform distribution of zero mean in interval $[-a, a]$. Fig.4 shows that the choice of a has influence on the prediction of ESN.

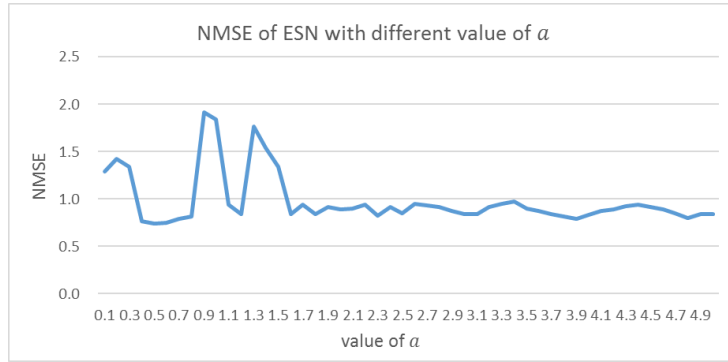


Fig. 4. NMSE of ESN with different value of a .

Considering these factors, we preprocess the weight matrix. The purpose of preprocessing is to keep the state of the reservoir away from the saturated area. Combined with Eq.(4), it means that W_{in} and W need meet the relationship in Eq.(8).

$$-\bar{s} \leq \sum_{n=1}^N W_{in}^n \cdot u(t) + \sum_{j=1}^N W_j \cdot x(t-1) \leq \bar{s}. \quad (8)$$

where s is the variables on the horizontal axis. W_{in}^n is the n_{th} row of W_{in} . W_j is the j_{th} row of W . According to the existing experiment, we set s as 2.65 in this paper. At this point, the derivative of the activation function is 2% of the maximum value.

Derived from the above formula, we can get the value of a [16].

$$a = \min_{t=1, \dots, p} a_t. \quad (9)$$

$$a_t \leq \bar{s} \cdot \sqrt{\frac{3}{(K+N) \cdot (\sum_{k=1}^K (u_k(t))^2 + \sum_{j=1}^N (x_j(t))^2)}}$$

where p is the length of input series. $u_k(t)$ is the k_{th} value of $u(t)$. $x_j(t)$ is the j_{th} value of $x(t)$. In the based ESN, a generally takes a value of 1. In our paper, we change this value for the purpose of a better prediction.

4.3 Activation function

Traditional ESN has a single type of memory function of reservoir. Considering the nonlinearity of the wavelet function, we replace some of the neurons in reservoir with the wavelet neurons [17]. The use of wavelet function in neural networks first appeared in the wavelet neural network (WNN). Wavelet theory has been widely used in numerical analysis and signal processing.

By enriching the activation function, the memory ability of the reservoir can be improved to a certain extent. There are many commonly used wavelet functions, including *Haar* wavelet, *Daubechies* wavelet and *Symlets* wavelet. In order to meet the definition of the activation region of neuronal activation functions, we choose the *Symlets* wavelet. In this way, the activation function presents a nonlinear behavior around zero.

Wavelet neurons have better spatial structure than ordinary neurons, which has been proved by a lot of experiments.

Eq.(10) is the definition of *Symlets* wavelet function [18].

$$\phi(x) = x \cdot e^{-\frac{1}{2}|x|^2} . \quad (10),$$

Fig.5 shows the *Symlets* wavelet function. It appears non-linearity in the activation area.

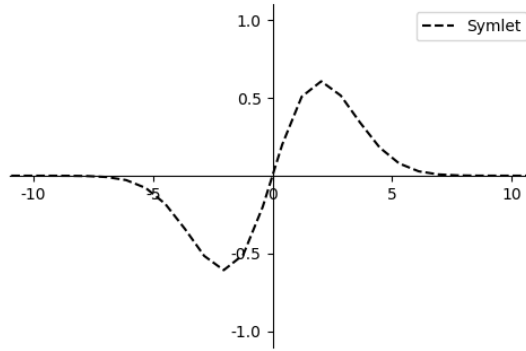


Fig. 5. *Symlets* wavelet function.

A wavelet mother function can transform many wavelet functions by translation and scaling. *Symlets* wavelet function after translating and scaling

$$\phi(x)_{d_j,t_j} = 2^{\frac{d_j}{2}} \cdot (2^{d_j} \cdot x - t_j) \cdot e^{-\frac{1}{2}(2^{d_j} \cdot x - t_j)^2} . \quad (11),$$

where $d_j = \frac{j}{N \cdot R_{mix}}$, $t_j = \frac{j}{N \cdot R_{mix}} - 0.5$, ($j = 1, 2, 3, \dots, N \cdot R_{mix}$, $N \cdot R_{mix} \in Z$). N is the size of reservoir. And R_{mix} represents the percentage of wavelet neurons to the whole neurons. In our paper, the percentage is 40%. By enriching the kinds of activation function and optimizing the initialization of weight matrix, the prediction ability of ESN is improved.

5 Result and conclusion

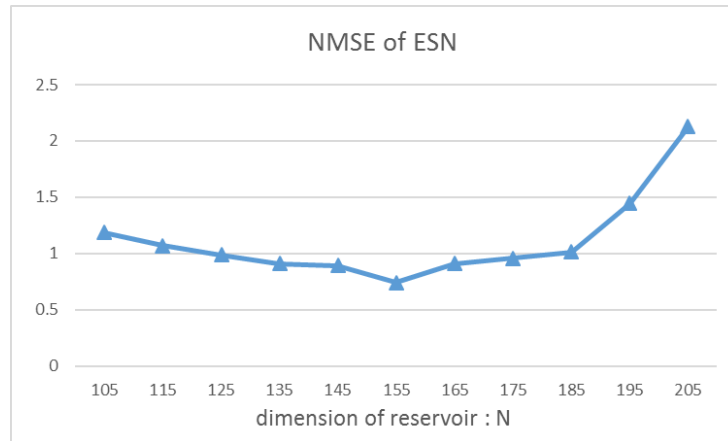


Fig. 6. NMSE of ESN with different dimension of reservoir.

Scale of dynamic reservoir may greatly affect the final prediction error. To compare the prediction effect of different networks, we calculate the NMSE of the ESN. Fig.6 shows the result of the basic ESN. As we can see in Fig.6, ESN gets the smallest NMSE when the dimension of reservoir is 155. And we would set $N = 155$ in the following experiments to test the predictive ability of the basic ESN and the improved ones.

We use WM-ESN to represent the ESN with optimized weight matrix. WF-ESN represents the ESN with mixed wavelet neurons. CM-ESN represents the ESN which enriches the kinds of activation function and optimizes the initialization of weight matrix. Table 1 shows the result of our experiments.

Table 1. NMSE of the network.

Model	Training time (s)	NMSE
ESN	0.479	0.7435
WM-ESN	1.323	0.6831
WF-ESN	1.542	0.6453
CM-ESN	2.545	0.5852

We compare two aspects of the networks. As our dataset is relatively small, we can not figure out the obvious difference in running time between these networks. It's certain that our improved CM-ESN has a more complex network structure and it may takes more time to train. This can be explained as we use higher model complexity in exchange for higher

prediction accuracy. Observing NMSE in Table 1, WM-ESN and WF-ESN both reduce the error to a certain extent and optimize the prediction effect. Based on this analysis, CM-ESN further reduces the NMSE. This means that the improved CM-ESN has a better ability to fit and learn our data, which is consistent with the previous analysis.

The purpose of our work is to predict the network traffic. Considering the non-linearity of traffic data, we choose echo state network (ESN). ESN is characterized by the simple structure and excellent ability of non-linear fitting. In this work, we first analyse the data from the telecom operator and get an overall judgment on the data. Then, we give an brief introduction on ESN. Considering the weakness of the basic ESN, we propose some improvement on selection of matrix initialization and activation function according to the existing literature. And we further improve the network by replacing part of the neurons with wavelet function. Result shows that our improvement is more effective compared to original ESN.

Acknowledgments. This work is partly supported by ZTE program “ The Prediction of Wireline Network Malfunction and Traffic based on Big Data”, the State Key Development Program of Basic Research of China (Grant No. 2013CB329005), the National Natural Science Foundation of China (Grant No.s 61571240, 61772287, 61601004, 61601005), the Priority Academic Program Development of Jiangsu Higher Education Institutions, the Natural Science Foundation of Jiangsu Province (Grant No. BK20161517), the Qing Lan Project, the Postdoctoral Science Foundation of China (Grant No.2017M611881), the National Engineering Research Center of Communications and Networking (Nanjing University of Posts and Telecommunications, Grant No.TXKY17003), the Scientific Research Foundation of NUPT (Grant No. NY217022).

References

- [1] China Internet Network Information Center, <http://www.cnnic.net.cn/hlwfzyj/hlwxzbg/hlwtjbg/201801/P020180131509544165973.pdf>, 2018/01/09
- [2] Chen, Y., Yang, B., Meng, Q.: Small-time scale network traffic prediction based on flexible neural tree. Elsevier Science Publishers B. V. (2012)
- [3] Pascale, A., Nicoli, M.: Adaptive Bayesian network for traffic flow prediction. *Statistical Signal Processing Workshop*. pp. 177–180 (2011)
- [4] Pang, Q., Cheng, S., Zhang, P.: Adaptive fuzzy traffic predictor and its applications in ATM networks. *IEEE International Conference on Communications*. pp. 1759–1763 (1998)
- [5] Lee K H. *First Course on Fuzzy Theory and Applications*. Springer Berlin, 27 (2008)
- [6] Li, S.: A predictability analysis of network traffic. Vol. 39(4), pp. 329-345. *Computer Networks* (2002)
- [7] Park, Y.M., Moon, U.C., Lee, K.Y.: A self-organizing fuzzy logic controller for dynamic systems using a fuzzy auto-regressive moving average (FARMA) model. Vol. 3, pp. 75-82. *IEEE Transactions on Fuzzy systems* (1995)
- [8] Gu, K., Sadiku, M.N.: Absorbing Markov Chain solution for Poisson's equation. *Southeastcon 2000. Proceedings of the IEEE*. pp. 297–300 (2000)
- [9] Wang, X., Shan, X.: A wavelet-based method to predict Internet traffic. *Communications, Circuits and Systems and West Sino Expositions, IEEE 2002 International Conference on*. pp. 690–694 (2002)
- [10] Han, M., Xu, M.: Predicting Multivariate Time Series Using Subspace Echo State Network. *Neural Processing Letters*. pp. 201–209 (2015)
- [11] Zhou, L., Wu, D., Dong, Z., Li, X.: When Collaboration Hugs Intelligence: Content Delivery over Ultra-Dense Networks. Vol. 55, pp. 91-95. *IEEE Communications Magazine* (2017)
- [12] Zhou, L., Wu, D., Chen, J., Dong, Z.: Greening the Smart Cities: Energy-Efficient Massive Content Delivery via D2D Communications. Vol. 14, pp. 1626-1634. *IEEE Transactions on Industrial Informatics* (2017)
- [13] Zhou, L., Wu, D., Chen, J., Dong, Z.: When Computation Hugs Intelligence: Content-Aware Data Processing for Industrial IoT, to appear in *IEEE Internet of Things Journal*, DOI: 10.1109/JIOT.2017.2785624.
- [14] Song, R., Xiao, W., Sun, C.: A new self-learning optimal control laws for a class of discrete-time nonlinear systems based on ESN architecture. Vol. 57, pp. 1-10. *Science China Information Sciences* (2014)
- [15] Xi, W., Hou, J., Shi, A., Zhang, X.: An improved forecast of ship sway based on ESN. *International Conference on Information, Cybernetics and Computational Social Systems*. pp. 422-425 (2017)
- [16] Wang, L., Qiao, J.F., Li, X.L.: A weight initialization method for echo state network. *Control and Decision*. pp. 356–360 (2016)
- [17] Zhang, J., Walter, G.G., Miao, Y., Lee, W.N.W.: Wavelet neural networks for function learning. Vol. 43, pp. 1485-1497. *IEEE Transactions on Signal Processing* (1995)
- [18] Chao, W., Qiang, G., Yuanlong, H., Runmin, H., Hao, M.: Adaptive complementary fuzzy self-recurrent wavelet neural network controller for the electric load simulator system. Vol. 8, pp. 1-12. *Advances in Mechanical Engineering* (2016)